Opgave 1 - Generel Bioinformatik (5%)

Du skal i opgaven skrive en sorteret liste af Blast e-værdier, med den mest signifikante (laveste) I toppen af listen.

Man skal huske at en e-værdi her er et udtryk for hvor mange Blast hits man vil forvente at finde med en given alignment-score eller bedre ved en tilfældighed. Det er altså et tal (ikke en sandsynlighed), et positivt tal, men ikke nødvendigvis et heltal og 0 er den lavest mulige e-værdi. Endvidere skrives en score f.eks som 1e-02, hvor 'e' er notationen som bruges for e-værdi eller Expect-value så med en e-værdi på 1e-02 menes 0.01 og du skal ikke forveksle 'e' med Exp på din lommeregner.

Svar:
C: 0
E: 1e-20
B: 1e-02
A: 0.02
D: 21

Opgave 2 – Metabolisme (20%)

Del 1.

Mht. valg af værktøjer og databaser gør jeg følgende overvejelser inden jeg går i gang:

- Mit udgangspunkt er en DNA sekvens fra en ukendt organisme.
- For at finde et svar på spørgsmålene (og kunne komme ud at købe foder til forsøgsorganismerne), vil jeg undersøge om der i de store brede sekvensdatabaser findes en homolog sekvens, hvortil der er knyttet information ang. oprindelsesart samt funktion af genets protein-produkt.
- Jeg vælger her BLAST som det mest oplagte valg.
- Jeg ved ikke om DNA sekvensen er protein-kodende, og vælger først at undersøge sagerne på DNA niveau, hvis jeg er heldig behøver jeg ikke at gå videre til protein-niveau.
- Jeg vælger at BLAST'e hos NCBI, da de har alle de store databaser tilrådihed (og det er den server vi har lært at bruge på kurset).
- Mere specifikt vælger jeg BLASTN mod den store brede database "NR/NT" (lige som vi har gjort i øvelserne).

Jeg får følgende BLAST hits:

```
Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           14,118,468 sequences; 36,366,575,838 total letters
Query= UnknownGene_CDS

Length=1500


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

ref|NM_001184352.1|  Saccharomyces cerevisiae S288c Subunit I ...  2706    0.0
ref|NM_001184357.1|  Saccharomyces cerevisiae S288c Endonuclea...  1754    0.0
emb|AM698041.1|  Vanderwaltozyma polyspora partial mitochondri...  1725    0.0
emb|FM995166.1|  Nakaseomyces bacillisporus complete mitochond...  1683    0.0
ref|NM_001184356.1|  Saccharomyces cerevisiae S288c I-SceII (A...  1303    0.0
gb|AE016821.1|  Ashbya gossypii (= Eremothecium gossypii) ATCC...  1258    0.0

....

>ref|NM_001184352.1| Saccharomyces cerevisiae S288c Subunit I of cytochrome c oxidase,
which is the terminal member of the mitochondrial inner
membrane electron transport chain; one of three mitochondrially-encoded
subunits (COX1), mRNA
Length=1605

 Score = 2706 bits (3000),  Expect = 0.0
 Identities = 1500/1500 (100%), Gaps = 0/1500 (0%)
 Strand=Plus/Plus
```

Ud fra mine BLAST hits (se ovenstående udpluk af oversigtstabellen), kan jeg se at jeg er så heldig at have fundet et perfekt hit (NM_001184352.1 – med en yderst signifikant e-værdig: så lille at den er afrundet til 0.0). Selve sekvensen findes altså i databasen, og jeg kan klikke på dette resultat og læse en masse detaljer om sekvensen.

I dette tilfælde får man faktisk alle svar forærende fra entry'ets titel:

```
LOCUS       NM_001184352           1605 bp    mRNA    linear   PLN 17-MAY-2010
DEFINITION  Saccharomyces cerevisiae S288c Subunit I of cytochrome c oxidase,
            which is the terminal member of the mitochondrial inner membrane
            electron transport chain; one of three mitochondrially-encoded
            subunits (COX1), mRNA.
```

Jeg vil i det følgende beskrive hvor man også kan finde denne information:

Del 1a):
Sekvensen stammer fra gær (Saccheromyces cerevisiae) – det kan man se fra SOURCE og ORGANISM linierne i GenBank entry'et:

```
SOURCE      mitochondrion Saccharomyces cerevisiae S288c
  ORGANISM  Saccharomyces cerevisiae S288c
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
```

Del 1b):

Der er tale om COX1:

```
     gene            1..1605
                     /gene="COX1"
                     /locus_tag="Q0045"
                     /gene_synonym="OXI3"
                     /db_xref="GeneID:854598"
     CDS             1..1605
                     /gene="COX1"
                     /locus_tag="Q0045"
                     /gene_synonym="OXI3"
                     /EC_number="1.9.3.1"
```

Del 1c):
Cox 1 koder for sub-unit 1 af cytochrome c oxidase:

```
     CDS             1..1605
                     /gene="COX1"
                     /locus_tag="Q0045"
                     /gene_synonym="OXI3"
                     /EC_number="1.9.3.1"
                     /note="Subunit I of cytochrome c oxidase, which is the
                     terminal member of the mitochondrial inner membrane
                     electron transport chain; one of three
                     mitochondrially-encoded subunits"
```

Del 1d):
Genet Cox1 findes i mitochondrie-genomet hos gær - det står flere steder, bla. i SOURCE (se Del 1a) og i noterne (se Del 1c).

BEMÆRK: Alternative måder (fx. BLAST på protein-nveau) at finde de korrekt svar på, giver fuldt point - det vigtigt er at dokumentation / argumentation er i orden.

Del 2.

Del 2e):
Der er flere måder at finde protein-sekvensen på. Det klart hurtigst er at tage protein sekvensen direkte fra det GenBank entry'et for Cox 1 vi fandt i del 1 (enten manuelt at klippe det ud, eller at følge protein-linket, og så bede om at se sekvensen i FASTA format).

```
                     /translation="MVQRWLYSTNAKDIAVLYFMLAIFSGMAGTAMSLIIRLELAAPG
                     SQYLHGNSQLFNVLVVGHAVLMIFFLVMPALIGGFGNYLLPLMIGATDTAFPRINNIA
                     FWVLPMGLVCLVTSTLVESGAGTGWTVYPPLSSIQAHSGPSVDLAIFALHLTSISSLL
                     GAINFIVTTLNMRTNGMTMHKLPLFVWSIFITAFLLLLSLPVLSAGITMLLLDRNFNT
                     SFFEVSGGGDPILYEHLFWFFGHPEVYILIIPGFGIISHVVSTYSKKPVFGEISMVYA
                     MASIGLLGFLVWSHHMYIVGLDADTRAYFTSATMIIAIPTGIKIFSWLATIHGGSIRL
                     ATPMLYAIAFLFLFTMGGLTGVALANASLDVAFHDTYYVVGHFHYVLSMGAIFSLFAG
                     YYYWSPQILGLNYNEKLAQIQFWLIFIGANVIFFPMHFLGINGMPRRIPDYPDAFAGW
                     NYVASIGSFIATLSLFLFIYILYDQLVNGLNNKVNNKSVIYNKAPDFVESNTIFNLNT
                     VKSSSIEFLLTSPPAVHSFNTPAVQS"
```

Alternativt kan man oversætte DNA sekvensen i Virtual Ribosome, her skal man huske følgende:

1) Man kan bruge ORF finderen og sætte krav om både START og STOP codon (vi ved at det er en komplet CDS).
2) Man skal huske at bruge den mitochondrielle genetiske kode.

Del 2f):

```
>COX1_Yeast
MVQRWLYSTNAKDIAVLYFMLAIFSGMAGTAMSLIIRLELAAPGSQYLHGNSQLFNVLVV
GHAVLMIFFLVMPALIGGFGNYLLPLMIGATDTAFPRINNIAFWVLPMGLVCLVTSTLVE
SGAGTGWTVYPPLSSIQAHSGPSVDLAIFALHLTSISSLLGAINFIVTTLNMRTNGMTMH
KLPLFVWSIFITAFLLLLSLPVLSAGITMLLLDRNFNTSFFEVSGGGDPILYEHLFWFFG
HPEVYILIIPGFGIISHVVSTYSKKPVFGEISMVYAMASIGLLGFLVWSHHMYIVGLDAD
TRAYFTSATMIIAIPTGIKIFSWLATIHGGSIRLATPMLYAIAFLFLFTMGGLTGVALAN
ASLDVAFHDTYYVVGHFHYVLSMGAIFSLFAGYYYWSPQILGLNYNEKLAQIQFWLIFIG
ANVIFFPMHFLGINGMPRRIPDYPDAFAGWNYVASIGSFIATLSLFLFIYILYDQLVNGL
NNKVNNKSVIYNKAPDFVESNTIFNLNTVKSSSIEFLLTSPPAVHSFNTPAVQS
```

Del 3.

Del 3g):
Som vi også har se gentagende gange gennem kurset, er der en større styrke i at sammenligne to sekvenser på protein-niveau, hvis de ikke er er super-ens (her vil DNA niveau være fint). Det skyldes følgende:

1) Bedre signal-til-støj forhold for sammenligninger på protein-niveau. DNA har kun fire "bogstaver" og proteiner har 20. Der vil være meget større sandsynlighed for at 2 DNA sekvenser er ens på en given position af stokastiske årsager (1/4 = 25%) end det er tilfældet på protein-niveau (1/20 = 5%).

2) Sammenligninger på protein-niveau bruger meget mere avancerede alignment-matricer end på DNA-niveau (fx. BLOSUM62), som indeholder afledt information om protein evolution.

3) BLASTP er i sin heuristik bedre egnet til at finde knap så gode hits end BLASTN: BLASTP har kun krav om 2 meget korte næsten-matches inden en sekvens udvælges fra databasen til fuldt local alignment - hos BLASTN er der krav om et perfekt-match på (default) 11bp.

Del 3h):
h1) Da jeg gerne vil arbejde med sammenligninger af protein-sekvenser vælger jeg BLASTP som værktøj (hos NCBI som før). Da vi KUN ønsker at sammenligne med humane sekvenser, har jeg en række muligheder for at indsnævre søgningen (**de er alle 100% korrekte**, men ID's på hits'ne kan varierer):

*) Vælge NR - og sætte organism feltet til Homo sapiens eller TaxID 9606.
*) Gå inde under de genom-specifikke databaser og find den humane (husk at vælge BLASTP so metode) - der er flere forskellige vej at finde frem til den humane database, man kan fx. følge det link til de genom-specifikke databaser, vi har bruger under BLAST øvelsen.

De mest signifikante hits når man søger i NR med "Homo sapiens" som filter, er som følger:

```
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           14,400,590 sequences; 4,930,896,956 total letters
Query= COX1_Yeast

Length=534



                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value
```

```
gb|ABU65184.1|  cytochrome oxidase subunit I [Homo sapiens]          607    3e-173
gb|ABR93038.1|  cytochrome c oxidase subunit I [Homo sapiens] ...    607    3e-173
gb|AEG23663.1|  cytochrome c oxidase subunit I [Homo sapiens]        607    4e-173
dbj|BAE14868.1|  cytochrome oxydase subunit I [Homo sapiens]         592    1e-168
gb|ABR10799.1|  cytochrome c oxidase subunit I [Homo sapiens]        591    2e-168
gb|ABR95529.1|  cytochrome c oxidase subunit I [Homo sapiens]        591    3e-168
gb|AAZ01660.1|  cytochrome c oxidase subunit I [Homo sapiens] ...    590    3e-168
gb|AAU02711.1|  cytochrome c oxidase subunit I [Homo sapiens] ...    590    3e-168
gb|AAP47921.1|  cytochrome c oxidase subunit I [Homo sapiens] ...    590    3e-168
gb|ABU46796.1|  cytochrome c oxidase subunit I [Homo sapiens]        590    4e-168
```

Bemærk: Alle hits'ne er i virkeligheden det samme gen - eftersom jeg har søgt i hele GenBank (med mere), kan jeg godt være ude for at mange varianter af det samme gen er blevet lagt ind.

h2) Der er adskillige hits med yderst signifikante e-værdiger. De bedste ligger med en e-værdi på 3e-173 (se ovenstående).

h3) De 2 øverste hits har præcis samme score, og jeg vælger her den ene (ABU65184.1) - alignment'et ser ud som følger:

```
>gb|ABU65184.1|  cytochrome oxidase subunit I [Homo sapiens]
Length=513

 Score =  607 bits (1566),  Expect = 3e-173, Method: Compositional matrix adjust.
 Identities = 300/529 (57%), Positives = 398/529 (75%), Gaps = 25/529 (5%)

Query  3    QRWLYSTNAKDIAVLYFMLAIFSGMAGTAMSLIIRLELAAPGSQYLHGNSQLFNVLVVGH  62
            RWL+STN KDI  LY +   ++G+ GTA+SL+IR EL  PG+  L GN  ++NV+V  H
Sbjct  4    DRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGN--LLGNDHIYNVIVTAH  61

Query  63   AVLMIFFLVMPALIGGFGNYLLPLMIGATDTAFPRINNIAFWVLPMGLVCLVTSTLVESG  122
            A +MIFF+VMP +IGGFGN+L+PLMIGA D AFPR+NN++FW+LP  L+ L+ S +VE+G
Sbjct  62   AFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLLASAMVEAG  121

Query  123  AGTGWTVYPPLSSIQAHSGPSVDLAIFALHLTSISSLLGAINFIVTTLNMRTNGMTMHKL  182
            AGTGWTVYPPL+   +H G SVDL IF+LHL  +SS+LGAINFI T +NM+   MT ++
Sbjct  122  AGTGWTVYPPLAGNYSHPGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQT  181

Query  183  PLFVWSIFITAFLLLLSLPVLSAGITMLLLDRNFNTSFFEVSGGGDPILYEHLFWFFGHP  242
            PLFVWS+ ITA LLLLS+PVL+AGITMLL DRN NT+FF+ +GGGDPILY+HLFWFFGHP
Sbjct  182  PLFVWSVLITAVLLLLSIPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGHP  241

Query  243  EVYILIIPGFGIISHVVSTYS-KKPVFGEISMVYAMASIGLLGFLVWSHHMYIVGLDADT  301
            EVYILI+PGFG+ISH+V+ YS KK  FG + MV+AM SIG LGF+VW+HHM+ VG+D DT
Sbjct  242  EVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVDT  301

Query  302  RAYFTSATMIIAIPTGIKIFSWLATIHGGSIRLATPMLYAIAFLFLFTMGGLTGVALANA  361
            RAYFTSATMIIAIPTG+K+FSWLAT+HG +++ +  +L+A+ F+FLFT+GGLTG+ LAN+
Sbjct  302  RAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTVGGLTGIVLANS  361

Query  362  SLDVAFHDTYYVVGHFHYVLSMGAIFSLFAGYYYWSPQILGLNYNEKLAQIQFWLIFIGA  421
            SLD+  HDTYYVV HFHYVLSMGA+F++  G+ +W P   G   ++  A+I F ++FIG
Sbjct  362  SLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAKIHFTIMFIGV  421

Query  422  NVIFFPMHFLGINGMPRRIPDYPDAFAGWNYVASIGSFIATLSLFLFIYILYDQLVNGLN  481
            N+ FFP HFLG++GMPRR  DYPDA+  WN ++S+GSFI+  ++ L I+++++
Sbjct  422  NLTFFPQHFLGLSGMPRRYSDYPDAYTTWNILSSVGSFISLTAVMLMIFMIWEAF-----  476

Query  482  NKVNNKSVIYNKAPDFVESNTIFNLNTVKSSSIEFLLTSPPAVHSFNTP  530
               + + V+  + P              S ++E+L   PP  H+F  P
Sbjct  477  --ASKRKVLMVEEP---------------SMNLEWLYGCPPPYHTFEEP  508
```

h4) Jeg kan se at jeg er på rette vej alene ud fra overskriften for ABU65184.1 i BLAST tabellen (cytochrome oxidase subunit I [Homo sapiens]), og jeg klikker ind på den for at læse mere.

Her kan jeg læse følgende detaljer, hvilket stemmer fint overens med den kendte funktion af gær-genet:

```
     Protein         1..513
                     /product="cytochrome oxidase subunit I"
     Region          10..495
                     /region_name="Cyt_c_Oxidase_I"
                     /note="Cytochrome C oxidase subunit I.  Cytochrome c
                     oxidase (CcO), the terminal oxidase in the respiratory
                     chains of eukaryotes and most bacteria, is a multi-chain
                     transmembrane protein located in the inner membrane of
                     mitochondria and the cell membrane of...; cd0166
```

Del 3i):

```
>gi|156456225|gb|ABU65184.1| cytochrome oxidase subunit I [Homo sapiens]
MFADRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGNLLGNDHIYNVIVTAHAFVMIFFMV
MPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLLASAMVEAGAGTGWTVYPPLAGNYSHPG
ASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQTPLFVWSVLITAVLLLLSIPVLAAGITMLL
TDRNLNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSI
GFLGFIVWAHHMFTVGMDVDTRAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTV
GGLTGIVLANSSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAKIHFTIMFIG
VNLTFFPQHFLGLSGMPRRYSDYPDAYTTWNILSSVGSFISLTAVMLMIFMIWEAFASKRKVLMVEEPSM
NLEWLYGCPPPYHTFEEPVYMKS
```

Opgave 3 (35%)

Del 1: Signalpeptidaserne fra to bakterier

a) Jeg brugte følgende søgestreng:
`organism:"Escherichia coli" AND name:"signal peptidase I"`
Da det skulle være fra UniProtKB/Swiss-Prot klikker jeg på "Show only reviewed" hvorefter søgestrengen er
`organism:"Escherichia coli" AND name:"signal peptidase 1" AND reviewed:yes`
og der er kun ét hit tilbage, med Accession nummer **P00803** og UniProt ID **LEP_ECOLI**.
Bemærk: hvis man søger med `organism:"Escherichia coli [562]"` hvilket man nemt kommer til, hvis man godtager UniProts forslag, kan man ikke finde det rigtige svar (man finder kun et fra UniProtKB/TrEMBL). 562 er TaxID for *E. coli* uden angivet stamme (*strain*); men alle dem med en angivet stamme hører også til *E. coli*! Herunder strain K12 med TaxID 83333. Derfor er det vigtigt at søge med organismenavn *uden* TaxID.

b) EC-nummer er **3.4.21.89**, gennavn er **lepB**, og sekvenslængden er **324**.

c) Næsten alle havde denne rigtig:

```
>sp|P00803|LEP_ECOLI Signal peptidase I OS=Escherichia coli (strain
K12) GN=lepB PE=1 SV=2
MANMFALILVIATLVTGILWCVDKFFFAPKRRERQAAAQAAAGDSLDKATLKKVAPKPGW
LETGASVFPVLAIVLIVRSFIYEPFQIPSGSMMPTLLIGDFILVEKFAYGIKDPIYQKTL
IETGHPKRGDIVVFKYPEDPKLDYIKRAVGLPGDKVTYDPVSKELTIQPGCSSGQACENA
LPVTYSNVEPSDFVQTFSRRNGGEATSGFFEVPKNETKENGIRLSERKETLGDVTHRILT
VPIAQDQVGMYYQQPGQQLATWIVPPGQYFMMGDNRDNSADSRYWGFVPEANLVGRATAI
WMSFDKQEGEWPTGLRLSRIGGIH
```

d) Dette løses nemmest ved at skrive "Bacillus subtilis" i Organism feltet på NCBI Protein BLAST søgesiden og vælge databasen "swissprot". Men får da et bedste hit med Accession nummer **P71013** og UniProt ID **LEPT_BACSU** og E-værdien **$5*10^{-13}$**.
Hvis man ikke har valgt databasen "swissprot" men i stedet "nr", bliver det mere kompliceret: man får 2 bedste hits, ZP_06874001.1 og NP_389324.1, begge med E-værdien **$2*10^{-12}$**,og ved at klikke på "9 more sequence titles" ved det andet hit finder man en UniProt-henvisning til **P71013** / **LEPT_BACSU**. Man kan også søge efter "NP_389324" i UniProt og dermed finde **P71013** / **LEPT_BACSU**. NB: Det giver ikke fuldt point at gennemføre søgningen uden at finde frem til UniProt accession og ID.
Spørgsmålet kan også løses ved at bruge BLAST-funktionen i UniProt, hvis man vælger "swissprot" som database. Resultaterne kan *efter* søgningen filtreres på taxonomi for at finde dem der er fra *B. subtilis*. Her får man dog et andet hit som det bedste, nemlig **P37943** / **LEPP_BACNA** med E-værdien **$3*10^{-9}$** (fra *B. subtilis* subsp. *natto*) – det er dermed også et korrekt svar.

e) Her ses alignmentet med P71013 / LEPT_BACSU fra NCBI Protein BLAST (NB: eftersom man allerede har et alignment i BLAST outputtet, er det overflødigt at lave et

nyt alignment):

```
> sp|P71013.1|LEPT_BACSU  RecName: Full=Signal peptidase I T; Short=SPase I; AltName:
Full=Leader peptidase I
Length=193

 Score = 68.6 bits (166),  Expect = 5e-13, Method: Compositional matrix adjust.
 Identities = 63/248 (25%), Positives = 102/248 (41%), Gaps = 80/248 (32%)

Query  60   WLETGASVFPVLAIVLIVRSFIYEPFQIPSGSMMPTLLIGDFILVEKFAYGIKDPIYQKT  119
            +LE G ++   + + L++R F++EP+ +   SM PTL  G+ + V             KT
Sbjct  20   YLEWGKAIVIAVLLALLIRHFLFEPYLVEGSSMYPTLHDGERLFV------------NKT  67

Query  120  LIETGHPKRGDIVVFKYPEDPKLDYIKRAVGLPGDKVTYDPVSKELTIQPGCSSGQACEN  179
            +   G  KRGDIV+    E  K+ Y+KR +G                 +PG  + Q  ++
Sbjct  68   VNYIGELKRGDIVIIN-GETSKIHYVKRLIG-----------------KPG-ETVQMKDD  108

Query  180  ALPVTYSNVEPSDFVQTFSRRNGGEATSGFFEVPKNETKENGIRLSERKETLGDVTHRIL  239
             L +                 NG +    +    K E ++ G+ L+      GD
Sbjct  109  TLYI-----------------NGKKVAEPYLSKNKKEAEKLGVSLT------GDFG----  141

Query  240  TVPIAQDQVGMYYQQPGQQLATWIVPPGQYFMMGDNRDNSADSRY-WGFVPEANLVGRAT  298
              P+                   VP G+YF+MGDNR NS DSR   G + E  +VG +
Sbjct  142  --PVK-------------------VPKGKYFVMGDNRLNSMDSRNGLGLIAEDRIVGTSK  180

Query  299  AIWMSFDK  306
             ++  F++
Sbjct  181  FVFFPFNE  188
```

Der er **25%** identiske positioner og **9** gaps (8 i databasesekvensen og 1 i query-sekvensen – tæl selv efter). Det længste gap er 19 positioner langt. NB: Vi har accepteret resultater, hvor der ikke er talt helt rigtigt, men det giver *ikke* fuldt point at påstå at der er 80 gaps. Der er 80 *positioner* med gap, men ét gap kan, som det fremgår, være op til 19 positioner langt.
Hvis man i stedet ser på alignmentet med P37943 / LEPP_BACNA fra UniProt BLAST, er der **26%** identiske positioner og **7** gaps (alle i databasesekvensen).

Del 2: Signalpeptiderne fra to bakterier

f) I Advanced search sætter man Field til Sequence annotation [FT] og Topic til Signal peptide. Kombineret med organisme-specifikationen giver det flg. søgestrenge:

```
organism:"Escherichia coli" AND annotation:(type:signal)
```

5224 hits

```
organism:"Bacillus subtilis" AND annotation:(type:signal)
```

319 hits

Hvis man i stedet har brugt `organism:"Escherichia coli [562]"` og `organism:"Bacillus subtilis [1423]"`, får man kun de hits der ikke har angivet nogen stamme eller underart, og det giver hhv. **299** og **314** hits. Dette er egentlig forkert, da alle stammer og underarter også hører til de to arter; men vi har givet fuldt point for dette svar, da det er så oplagt at komme til at lave det på denne måde.
Det er derimod forkert at skrive `"Escherichia coli"` uden `organism:` foran. Det giver nogle ekstra hits der omtaler E. coli et eller andet sted i entry'et uden at være fra E. coli – heriblandt et humant entry!

g) Der er kun én rigtig måde at løse dette på: Når man i Advanced search har valgt Field

til Sequence annotation [FT] og Topic til Signal peptide, skal man også sætte Confidence til Experimental. Det giver flg. søgestrenge:

```
organism:"Escherichia coli" AND annotation:(type:signal AND
confidence:experimental)
```

232 hits

```
organism:"Bacillus subtilis" AND annotation:(type:signal AND
confidence:experimental)
```

43 hits

Med de alternative organismespecifikationer (562 og 1423) giver det hhv. **232** og **43** hits. Det er ikke nok at tilføje `reviewed:yes`. Dette er kun en tilkendegivelse af, at der har været en manuel bedømmelse af entry'et (herunder hvorvidt evidensen er eksperimentel eller ej).
Det er heller ikke korrekt at bruge `existence:"evidence at protein level"` – det fortæller at der er eksperimentel evidens for proteinet, men ikke nødvendigvis for signalpeptidet.

h) Man sorterer ved at klikke på de små pile ud for Entry name. Hvis man har lavet de foregående spørgsmål helt korrekt, bliver listen:
ACRA_ECOLI, AG43_ECOLI, AGP_ECOLI, AIDA_ECOLX, ALSB_ECOLI
Andre svar giver også fuldt point, hvis de er konsistente med svarene på de foregående spørgsmål.

i) Her er det vigtigt at tage sammenhængende vinduer af sekvens – de 15 sidste af signalpeptidet efterfølges umiddelbart af de første 5 af det færdige protein. Hvis man har lavet de foregående spørgsmål helt korrekt, bliver det:

```
LAVVLMLSGSLALTGCDDKQ
VALSLAAVTSLPVLAADIVV
AAVAGIVLLASNAQAQTVPE
LLVLAVVSTIGNAFAVNISG
SGTLVGLMLSTSAFAAAEYA
```

Igen kan andre svar give fuldt point, hvis de er konsistente med svarene på de foregående spørgsmål.

j) Position **15** er den mest konserverede, her er 9 A'er ud af 10 (i hvert fald i vores eget svar på spørgsmål i). Der er 5 A'er ud af 5 hvis man blot bruger sekvenserne givet i opgaveteksten. Dette er den sidste position **i signalpeptidet.**

Del 3: Forudsigelse af signalpeptider

k)
Ved brug af vores svar på spørgsmål i:

$f(L) = 2/5 = 0.4$

$g(L) = f(L)*q(L|L) + f(V)*q(L|V) + f(A)*q(L|A) + f(S)*q(L|S)$
$= 0.4*0.38 + 0.2*0.13 + 0.2*0.06 + 0.2*0.04 = 0.198$

$p(L) = (\alpha*f(L) + \beta*g(L)) / (\alpha+\beta) = (4*0.4 + 10*0.198) / (10+4) = 0.2557$

w(L) = 2*log(p(L)/q(L))/log(2) = 2*log(0.2557/0.099)/log(2) = **2.738**

Ved brug af sekvenserne givet i opgaveteksten (reserveløsning):
f(L) = 2/5 = 0.4
g(L) = f(L)*q(L|L) + f(V)*q(L|V) + f(A)*q(L|A) + f(I)*q(L|I)
= 0.4*0.38 + 0.2*0.13 + 0.2*0.06 + 0.2*0.17 = 0.224
p(L) = (α*f(L) + β*g(L)) / (α+β) = (4*0.4 + 10*0.224) / (10+4) = 0.2743
w(L) = 2*log(p(L)/q(L))/log(2) = 2*log(0.2743/0.099)/log(2) = **2.941**

Der er i en del tilfælde givet delvis point, hvis formlerne har været rigtige men udregningen forkert (f.eks. forkerte opslag i tabellen).

l) Positionerne **15** og **13** indeholder mest information, og **Alanin** er den hyppigste aminosyre på begge disse positioner. Logoet er lavet af **174** sekvenser.

m) Pearson korrelationskoefficient: **0.35616**
Aroc værdi: **0.99106**

n) Her er et uddrag af outputtet fra EasyPred:

```
Number Sequence Assignment Prediction


1      MKQSTIALALLPLLFTPVTK    0.000   -8.429
2      KQSTIALALLPLLFTPVTKA    0.000  -12.767
3      QSTIALALLPLLFTPVTKAR    0.000   -5.407
4      STIALALLPLLFTPVTKART    0.000  -11.192
5      TIALALLPLLFTPVTKARTP    0.000    1.354
6      IALALLPLLFTPVTKARTPE    0.000   14.333
7      ALALLPLLFTPVTKARTPEM    1.000   12.946
8      LALLPLLFTPVTKARTPEMP    0.000    2.144
9      ALLPLLFTPVTKARTPEMPV    0.000    0.775
10     LLPLLFTPVTKARTPEMPVL    0.000  -10.775
```

Heraf ses, at scoren for det rigtige vindue (Assignment=1) er **12.946**
Vinduet umiddelbart før dette scorer faktisk højere, nemlig **14.333**

o) Med *B. subtilis* testvinduerne fås:
Pearson korrelationskoefficient: **0.27160**
Aroc værdi: **0.93572**
Dvs. at den er **dårligere** end når man tester på *E. coli*. Der er åbenbart så store forskelle i signalpeptidernes udseende mellem Gram-positive og Gram-negative bakterier, at man ikke uden at ofre en hel del i performance kan bruge en metode trænet på *E. coli* til at forudsige på *B. subtilis*.

Opgave 4 – Protein drugs (25%)

Du er blevet kontaktet af en firma, der arbejder på at udvikle protein drugs. Firmaet arbejde på at udvikle et produkt på basis følgende protein sekvens:

```
>QUERY
KKASTIFGMPLQQDPVPATSTFIVSDFLQFLQTAVTCFNKLRIPEERFPLYLAGVFPNC
PETQCFVRCLSANLNLYCDETGSDIDRHYLQYGLGQDYNCFRQKAEQCLAANTSPCNDP
CEAAYKQELCFLDEFRKYVDSNMNSLIAAVAVEKAEQNPVYYNMLAHN
```

Del 1: homologi model

a) Firmaet vil have dig til at lave en homologi-baseret model af proteinets 3 dimensionelle struktur. Hvilke af følgende 3 protein strukturer vil du vælge til at lave homologi modellen

1: 3L47 chain A
2: 1PLC chain A
3: 2X47 chain A

Begrund dit svar!

Jeg submitter min query til Blast, aligner mod PDB og finder at PDB entry 3L47 chain A er et signifikant hit med en e-value på $2\ 10^{-8}$. Svaret er derfor 1

b) Find 4 cysteiner (C) ud fra listen nedenfor (numrene svarer til position og aminosyrer i proteinsekvensen), der med stor sandsynlighed kan danne disulfidbindinger (disulfidbroer) in denne proteinsekvens. Disulfidbindinger er to cysteine aminosyrer, der sidder i tæt fysisk kontakt og danner en kovalent binding. Disse bindinger er generelt stærkt konserveret indenfor en proteinfamilie.

59 C
64 C
68 C
77 C
99 C
107 C
115 C
119 C
128 C

Jeg kigger på den alignment Blast laver af min query mod 3L47.A og finder

```
query  51   YLAGVFPNCPETQCFVRCLSANLNLYCDETGSDIDRHYLQYGLGQDYN-CFRQKAEQCL-  108
            Y A  FP+ P T CFVRC+   LNLY D+ G D+  ++   G   D +  F  K   CL
Sbjct  27   YRANEFPDDPVTHCFVRCIGLELNLYDDKYGVDLQANWENLGNSDDADEEFVAKHRACLE  86
```

```
Query  109  AANTSPCNDPCEAAYKQELCFLDEFRKYVDSN  140
            A N     D CE AY    C  +++  Y ++N
Sbjct  87   AKNLETIEDLCERAYSAFQCLREDYEMYQNNN  118
```

Her ser jer at 64C, 68C, 107C, 119C og 128C er bevaret mellem de to proteiner. Jeg kan ikke udfra alignmentet afgøre hvilke af disse 5 der kan indgå i disulfidbindinger

c) Angiv, hvordan de 4 cysteiner med stor sandsynlighed danner par (f.eks 59 med 128, og 64 med 107. Bemærk disse par er IKKE det rigtige svar). Begrund dit svar.

Jeg kigger på strukturen af 3L47 og finder de 5 cysteiner fra spoergsmål b udfra blast alignmenten. På strukturen finder jeg at C107 – C128, og C64 – C119 sidder i tæt kontakt og dermed med stor sandsynlighed kan danne par. C68 indgår med stor sandsynlighed IKKE i en disulfidbinding.

Hint: For at finde de 4 cysteiner i den PDB struktur du har valgt til at repræsentere din query sekvens, skal du kigge på BLAST-alignmentet og matche dine query-aminosyrer til aminosyrerne i PDB-strukturen.

(opgaven fortsætter på næste side)

Del 2: Peptid-binding

Du har modtaget følgende peptid-data fra et laboratorium

YTDKIAMSY
YANMWSLMY
VTDTALAYF
TSASFTDLY
SVDSDHLGY
RSDEYVAYY
NTDNKFISY
NSDEQSLEY
MTDVDLNYY
MTDKICWLY
MTAASYARY

Peptiderne er alle målt til at binde til receptoren X. Firmaet forsøger at udvikle peptid-drugs, der kan binde til receptoren X og dermed hæmme dens effekt. Du kan med fordel benytte EasyPred til at besvare nedenstående spørgsmål.

d) Kan du udfra de 10 peptider angive hvilke peptid-positioner, der har mest betydning for bindingen til receptoren X? Begrund dit svar.

Jeg benytter EasyPred og laver et sekvens logo udfra de 10 peptider med defaults settings. Udfra logoet finder jeg at P9, P3 og P2 har det største informations indhold, og dermed største betydning for binding

e) Firmaet har tre mulige peptider, de vil teste for binding til receptoren X. Kan du udfra de 10 peptider vist overfor, angive hvilket af de 3 peptider nedenfor, der vil have størst chance for at binde til receptoren? Begrund dit svar

1: VTDEGTSSF
2: TLDSEDGLY
3: EYKLQQGTF

Jeg benytter igen EasyPred med default settings, og scorer de tre peptider mod vægt matricen lavet udfra de 10 peptider fra spørgsmaal d). Her finder jeg

1 VTDEGTSSF 13.459
2 TLDSEDGLY 9.461
3 EYKLQQGTF -5.492

og dermed at VTDEGTSSF vil have den største chance for at bind til recpetoren

f) Det viser sig, at peptider med F eller Y i C-terminalen (den sidste position) har stærke side-effekter, og derfor ikke kan bruges som drug. Hvilken af følgende aminosyrer vil du foreslå firmaet at ændre den sidste aminosyre til således at peptidet fortsat vil binde, men ikke have de nævnte side-effekter forårsaget af Y og F? Begrund dit svar

1: L
2: Q
3: D

Jeg benytter atter EasyPred og submitter scorer peptiderne VTDEGTSSL, VTDEGTSSQ, VTDEGTSSD mod matricen lavet udfra de 10 peptider fra spørgsmaal d) Her finder jeg at peptidet med L på C terminalen har den højeste score og dermed vil jeg vælge L.

Opgave 5 - Fylogeni (15%)

Der er mange programmer og webservere som kan benyttes til at lave multiple alignment. Ingen af dem er decideret forkerte, men nogen er klart bedre end andre. I løbet af undervisningen har I specielt hørt om 2 gode bud: (1) blandt multiple aligment serverne på EBI, viste "mafft" sig i en øvelse at være langt bedre end "clustalw" og lidt bedre end "muscle". (2) Eftersom de sekvenser i skal aligne koder for protein, kan det betale sig at inddrage aminosyre-niveauet i alignmentet - RevTrans serveren på CBS gør netop det.

Der er også flere måder at lave et neighbor joining træ på. De to mest oplagte: (1) Vha. TreeHugger serveren på CBS, som I har benyttet i en øvelse: Man giver serveren et alignment, enten ved at paste eller ved at uploade en fil, og klikker derefter på submit. (2) Vha ClustalX multiple alignment programmet (den grafisk klient som kan downloades til ens egen computer): man åbner sit multiple alignment i ClustalX, og vælger derefter menuen "Trees" og derefter "Draw Tree". Resultatet er i begge tilfælde en fil i det såkaldte Newick format, hvor træet er repræsenteret vha parenteser, kommaer, og navne på blade:

(Rat:0.086218,Mouse:0.083654,(((Carp:0.106593,Xenopus:0.120971):0.014647,Chicken:0.117083):0.008674,((Whale:0.107505,(Bovine:0.085296,Seal:0.099960):0.004033):0.005629,Human:0.118089):0.003946):0.020172);

Newick filen åbnes i FigTree, og man vælger “Carp” som outgroup ved at (1) klikke på den gren der fører op til “Carp”, (2) klikke på “Reroot” knappen på FigTree toolbar’en:

c1) Den nærmeste slægtning til koen (“Bovine”) i træet her er sæl (“Seal”): Den direkte forfader til koen (angivet med en ring på figuren) er også forfader til sælen (her i træet – der er arter vi ikke har med!).

c2) Mennesket er i det her træ nærmere beslægtet med hvalen end med musen: Man skal kun gå en forfader tilbage fra mennesket for at finde en som også har hvalen som efterkommer (angivet med en ring i figuren). Man skal gå to led tilbage fra mennesket for at finde en som også har musen som efterkommer.

c3) Xenopus (en slags frø) har udgrening tættest på roden (bortset fra outgroupen).